DOCUMENT RESUME

ED 466 693 TM 034 272

AUTHOR	Hickey, Daniel T.; DeCuir, Jessica; Hand, Bryon; Kyser, Brandon; Laprocina, Simona; Mordica, Joy
TITLE	Technology-Supported Formative and Summative Assessment of Collaborative Scientific Inquiry.
SPONS AGENCY	National Science Foundation, Arlington, VA.
PUB DATE	2002-04-00
NOTE	13p.; Supported by a Seed Grant from the Center for Innovative Learning Technologies. Project collaborators include University of Georgia, Atlanta; Georgia Tech; Georgia State University; Ogden County Schools; Virginia Tech; King's College; NASA Classroom of the Future; WestEd; SRI International; and University of Washington.
PUB TYPE	Reports - Research (143)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	Accountability; *Cooperative Learning; *Evaluation methods; Group Instruction; *Middle School Students; Middle Schools; Scoring Rubrics; Student Evaluation; *Student Participation; Video Equipment; *Videotape Recordings

ABSTRACT

This study defined and validated a new set of dimensions, new anchoring descriptions, and a new rubric format for assessing participation in collaboration. One strand of the research explored the use of analog video-technology to conduct summative assessment of collaborative inquiry. The second strand of the research explored the use of video digital technology to conduct formative assessment of collaborative inquiry. Participants were from seven middle school classrooms taught by two teachers at two schools. Students in all classrooms were asked to complete a brief genetics performance assessment, and 42 student triads were videotaped for the 2 research strands. In the first strand, five graduate students evaluated the collaborations. It appeared that the summative assessment practice attained a level of precision sufficient for comparing groups of students to each other, although it did not appear that this approach is likely to yield the precision needed by any formal accountability system. In the second strand, triads of students (15 sets) were asked to review their own assessment tapes and then score their own collaboration using a scale developed for the purpose. This approach does appear to be a promising method of enhancing participation in collaboration and increasing students' ability to engage in collaborative learning that might be implemented on a larger scale.(Contains 4 tables and 29 references.) (SLD)

Technology-Supported Formative and Summative Assessment of Collaborative Scientific Inquiry

Daniel T. Hickey
Jessica DeCuir
Bryon Hand
Brandon Kyser
Simona Laprocina
Joy Mordica

Learning & Performance Support Laboratory
University of Georgia[1]

This project used diverse views to define and validate ways to assess and promote collaboration and participation in collaborative inquiry in science. The value of such an effort is illustrated by the many educational standards documents that highlight the value of the ability to work with others (e.g., Kendall & Marzano, 1997). Haertel and Means' (2000) review of current educational research methodology highlighted the need for common instruments & methods for assessing valued learning outcomes. The ascendance of standards-oriented reforms highlights the value of characterizing outcomes in a manner that can be readily communicated to diverse stakeholders while simultaneously directing useful educative activity. Haertel & Means also highlighted the need for new measures that reflect contemporary views on knowing and learning, particularly for documenting outcomes promised by technology-supported innovations. While all educational standards acknowledge the importance of collaboration, this elusive construct confounds conventional assessment and evaluation practices. Particularly in the current accountability-oriented climate, educational outcomes that cannot be readily measured and communicated to diverse stakeholders are likely to be overlooked.

This effort built on the PI's prior attempts to evaluate the impact of GA Tech's *Learning by Design* middle school science curriculum (Kolodner et al, 1998, in review). In one part of this evaluation, over 150 teams in LBD and comparison classrooms were videotaped while collaboratively completing a performance assessment obtained from the *Performance Assessment Links in Science* (PALS, 2001) website. These videotapes were scored using nine dimensions of collaboration advanced in prior research by Pomplun (1996). These scores, along with the scores of the students' performance assessments, were analyzed and reported as evidence of the effectiveness of the LBD curriculum (Hickey, 1999; 2000b; an extended discussion of the theoretical and practical issues in this approach is presented in Hickey, 2001)

[1] The project was supported by a Seed Grant from the Center for Innovative Learning Technologies, which is supported by the National Science Foundation. We gratefully acknowledge the participation of the students and teachers in the two classroom where this research was conducted. We also acknowledge the participation of the following collaborators on this effort: Simona Laprocina, Paula Schwanenflugel, Elizabeth Meisinger (University of Georgia) Jennifer Holbrook & Jackie Gray (GA Tech); Joy Mordica, Nancy Schafer (GA State), Juan Balderas & Vicki Winn, Ogden County Schools; Dan Dunlap, VA Tech; Richard A. Duschl, King's College, London; Sanna Järvelä, University of Oulu, Finland; Steven McGee, NASA Classroom of the Future; Sharon Nelson-Barber, WestEd; Bill Penuel, SRI International; and Steven Tanimoto, University of Washington. For more information, contact Daniel T. Hickey, 611 Aderhold Hall, Athens, GA 30602, dhickey@coe.uga.edu.

The present project extended and refined this assessment practice. Specifically, we defined and validated a new set of dimensions, new anchoring descriptions, and a new rubric format for assessing participation in collaboration, and attempting to use these to help learners evaluate and improve their own collaboration. Thus, one strand of our research explored the use of analog video-technology to conduct summative assessment of collaborative inquiry. The second strand of our research explored the use of digital video technology to conduct formative assessment of collaborative inquiry. The combination of these two strands in a single effort directly addresses contemporary concerns with the relationship between validity and value among educational researchers. Following from Frederiksen and Collins (1989), we explore the relationship between evidential and consequential validity (Hickey, Wolfe, & Kindfield, 2000) relative to collaborative inquiry. Specifically we considered whether a marginally reliable summative assessment of collaborative inquiry can still be valid because it supports learning and communicates value to learners, educators, and policy makers.

Our efforts build on contemporary views of learning (Bransford, Brown, & Cocking, 1999) and formative assessment (e.g., Black & Wiliam, 1998, Graue, 1993, Turnstall & Gipps, 1996). We are searching for modest, scaleable assessment practices that motivate learners to engage in effective "assessment conversations" (Duschl & Gitomer, 1997) that promise to dramatically enhance student learning. This activity is characterized by authentic scientific argumentation (e.g., Driver, Newton, & Osborne, 2000) in which students are making and warranting knowledge claims based on evidence and theory (e.g., Jiménez-Aleixandre, Rodríguez, & Duschl, 2000).

Along the way we grappled with some of the issues that contemporary sociocultural views of knowing and learning (e.g., Greeno, 1998; Wenger, 1998) present when considering what it means to define, assess, and promote collaboration and collaborative learning. Divergent perspectives on learning yield remarkably divergent conceptualizations of collaboration. Our use of *collaboration & participation* acknowledges the tension between the conventional notion of individuals acquiring domain-general skills & dispositions that support collaboration, and contemporary sociocultural notions of shared inquiry-oriented practices becoming ritualized within particular communities of learners. A fundamental assumption of this project is that a single definition of "collaboration" that is broadly meaningful to learners, educators, researchers, and policy makers will be valuable for fostering collaboration and the ability to collaborate in our schools. The project was deliberately set up to present the conflicts and contradictions that need to be negotiated in order to meet this admittedly idealistic long-term goal.

METHOD

The participants in this study were from seven classrooms taught by two teachers at two schools that previously participated in the *GenScope Assessment Project* (Hickey, 2000a). These schools were in the same school district but served very different student populations. One school served a relatively lower-SES suburban community. Over 30% of the students at this school had qualified for the federal lunch subsidy, and nearly every student (99.5%) was African American. The school typically posted average achievement scores that were below the national average, but higher than most of the other schools in this district that also served predominantly African American students. State data showed that 61% of these students passed the science component of the high school graduation test on their first attempt. The other was a high SES suburban school where 12% of the students were non-white, 1.5% of the students received subsidized lunch, and 95% of the students passed the science graduation test on their first attempt. Data from the GenScope Assessment Project showed substantial disparity in genetics knowledge between the two sets of classrooms. As in the previous GenScope evaluations (e.g., Hickey, Kindfield, Horwitz, & Christie, 1999), mean proficiency in the lower-SES classrooms *after* genetics instruction was lower

than the mean performance in the higher SES classrooms *before* instruction. Specifically, performance on conventional genetics content measures and performance assessments a few weeks prior to the first data collection in this study differed by at least 1.5 SD. In homogenous populations in our research, this difference equals roughly 2 grade levels.

Students in all seven classrooms were asked to complete a brief (about 30 minute) genetics performance assessment (called *Human Inheritance*) obtained from the PALS Website.. Forty-two of these triads were videotaped (6 per classroom) with 8mm cameras equipped with wide-angle lenses and tabletop microphones. These tapes were then used in two different strands of research, described next.

Summative Assessment of Collaborative Inquiry

The first strand of this effort was refining a set of dimensions and corresponding practice for assessing collaborative inquiry. Five graduate students used ten of the videotapes of collaboration to interpretively and empirically analyze 29 candidate dimensions following from Pomplun (1996), Gray (2001), and Jimenez-Aleixandre, Rodriguez, & Duschl (2000). They were asked to identify a small subset that reflected the perceived needs of researchers, educators, and policy makers. They were specifically instructed to exclude dimensions that presented concerns about equity or cultural bias, and to include dimensions that reflected contemporary situative perspectives that emphasize how domain knowledge is constructed within and partially bound to ritualized collaborative inquiry.

These five students spent roughly 20 hours working together during five separate meetings. This effort was further informed by a half-day workshop that occurred about halfway through the process, conducted at AERA 2001. The PI and the five graduate students along with ten outside researchers (Dunlap, Duschl, Gray, Holbrook, Jarvela, Nelson-Barber, McGee, Penuel, Ravitz, Tanimoto) discussed the various dimensions, equity issues, measurement issues, and policy issues. This meeting lent general support to the importance of the effort, the selected dimensions, and the validity of the approach, and laid the groundwork for subsequent collaboration. The effort of the five graduate students resulted in the six dimensions shown on Table 1.

In order to assess whether or not it was possible to score the six final dimensions reliably, three of graduate students (Hand, Kyser, & Laprocina) independently scored 23 of the remaining videotapes according to the six dimensions on Table 1. Scorer 1 scored all 23 of the tapes while scorers 2 and 3 scored 10 and 13 tapes, respectively. As in the earlier effort using the Pomplun dimensions, inter-rater reliabilities were disappointingly low. The correlation of the summed scores between the first and second raters on 13 tapes was only .74, while the correlation of the summed scores between the first and third raters on 10 tapes was .56. Examining the correlations for each of the six scales showed a mixed pattern of divergence across the scales, with scorer 1 and scorer 2 diverging the most on Scale 3, and scorer 1 and scorer 3 diverging most on Scale 4 and Scale 5. Particularly given that these correlations do not include a correction for chance (i.e., Cohen's K), our observed reliabilities are problematic from conventional measurement perspective. Of course, correlations based on such small numbers are fickle, and further training of scorers is likely to increase inter-rater reliability. Nonetheless, the present effort represented a substantial, coordinated effort on the part of motivated, thoughtful graduate students. Coupled with the modest reliabilities obtained previously, it seems that reliable summative assessment of collaborative inquiry via the present method will continue to be problematic.

Equity. One of the issues we struggled with throughout this process concerned equity. We were concerned that culturally specific styles of interaction might bias scores. Specifically, we were concerned that interaction styles that are culturally appropriate in a racial minority community might be viewed as disrespectful or uncooperative by observers from a mainstream culture. We discussed this issue at length while defining the dimensions, and relied a great deal on the input

from the two African American participants (Joy Mordica, who completed a term paper on the topic for a course on the psychology of inner city learners taught by Asa Hilliard at GSU, and Jessica DeCuir, whose UGA Ed Psych dissertation topic is African American Identity). Additional insights were provided at the AERA meeting by Sharon Nelson-Barber, a leading expert in the area of cultural validity and equity in assessment.

A preliminary review of the videotapes scored in the summative assessment revealed interesting examples of African American students effectively using African American Vernacular English (AAVE), a dialect of English, to successfully negotiate shared understanding of content and collaboration standards. It seems likely that many scorers who are unfamiliar with AAVE would be inclined to give these examples poor marks on some dimensions (e.g., *Collaborative Participation*) because of their difficulty discerning this dialect or possibly because of racial biases. Our expectation is that including the *Warranting Knowledge Claims with Data* and *Warranting Knowledge Claims wit Theory* dimensions would offset this problem by both giving scorers another dimension that would be scored more highly, and by reminding scorers to try to "look past" cultural biases on the other more domain-general dimensions. We had hoped to explore these issues in more interpretive and empirical detail. However, neither of the African American graduate students who participated in the original scale definition participated in the subsequent scoring of the 23 additional tapes, and the other were not familiar enough with AAVE or the cultural norms of African American secondary students to represent that perspective in the effort.

Summary. In summary, we believe that our summative assessment practice attained a level of precision sufficient for comparing groups of students to each other. For example, it seems that this practice is a valid and appropriate way of comparing students in two different curricular environments in the same domain (as in the LBD program evaluation). This approach appears equally appropriate for a wide range of students who might be assessed in western nations. Additional work is needed to reduce unexplained variance before making more precise interpretations, such as the differences between two teachers implementing the same curriculum. In particular, it seems that more reliable scores would have been obtained more quickly had we spent more time identifying benchmark examples to train scorers. Better yet, we might consider having students dramatize examples to clearly illustrate what different levels of each of the dimensions look like.

These improvement aside, the results of this project and the prior effort suggests that this approach to summative assessment of collaboration is unlikely to ever yield the precision demanded by any formal accountability system. Indeed, as a *purely* summative assessment, it seems that our model of assessment practice actually has rather limited value.

Formative Assessment of Collaborative Inquiry

The second strand of our effort follows from Fredriksen and Collins' (1989) notion of "systemic validity":

> A systemically valid test is one that induces in the educational system curricular and instructional changes that foster the development of the cognitive skills that the test is designed to measure. Evidence for systemic validity would be an improvement in those skills after the test has been in place within the educational system for a period of time (p 27).

From this perspective, a systemically valid collaboration assessment practice *directly* enhances student collaboration; this contrasts with the conventional view that assessment indirectly supports learning by identifying more or less effective programs. A systemically valid collaboration

assessment rubric must be meaningful and valued by both educators and students (as well as by researchers and policy makers); systemically valid assessment practices must be readily implementable in a manner that allows learners to reflect on and improve their collaborative participation.

Twenty-five of the video-taped examples of collaborative inquiry were digitized, compressed (using the *RealMedia* format), and placed on laptop computers. A team returned to the same classrooms where they were recorded in order to explore a simple, scalable model of formative assessment practice for collaborative inquiry. We first asked each traid to review their completed performance assessment using a modified version of the scoring rubric provided with the assessment. We then asked each triad to view videos of themselves or their classmates (for the triads who were not recorded originally) and to score their own collaboration using the scale shown on Table 1. Both of these sessions were followed by a feedback session where groups were asked to volunteer the various scores they assigned and to defend their answers. In both sets of classrooms, the first author began the feedback session and invited the teachers to take over when the felt comfortable doing so (which generally occurred during the first class period). Videotape recordings were made of the researcher/teacher and 3-4 triads in each of the classrooms.

Recollections of our experience and videotape of this activity were analyzed to consider whether these practices (1) can be readily implemented in science classrooms; (2) appear to support student learning of specific science content; (3) appear to support student learning of and value for collaborative inquiry, and (4) are effective in diverse classrooms.

Scalability. With regard to the first research question, both classrooms were able to readily appropriate the model of practice. Videotapes from the group activities revealed spirited discourse and argumentation around both the performance assessment and the collaboration standards. We were particularly encouraged to see how some of the groups' independent efforts to clarify the distinction between the six dimensions of collaboration scaffolded their participation in a whole class discussion. We were also encouraged by the ease with which the teachers appropriated the entire activity after the PI's initial demonstration. Given that a previous concerted effort to support "assessment conversation" (Duschl & Gitomer, 1997) in these classrooms had been rather unsuccessful (Hickey, Kruger, Fredrick, Schafer, & Kindfield, 2002), we found this initial success quite encouraging.

Content learning. Regarding the second research question, the videotape recordings of the groups scoring their performance assessment and the whole class feedback session provide clear evidence of student learning. The videotapes show numerous examples of successful negotiation of the shared meaning of the scoring rubric within the collaborative groups. We were also impressed by the ease with which these teachers maintained a fairly high level of engaged discussion around the targeted content when reviewing group scores. It certainly appeared that students understanding of the targeted concept as presented in the rubric increased substantially across both activities. Particularly heartening was the way groups adjusted their scores as other groups read and defended their answers. Students appeared to attach genuine value to the domain knowledge.[2]

That students appeared to learn science content via this activity is not a noteworthy finding, given the extensive body of literature showing just that. However, we were particularly encouraged by the cases where the group had relied on the insights of just one knowledgeable individual to provide the correct answer when completing the assessment, sometimes without input from the

[2] In order to provide empirical individual-level evidence of increased content understanding, we also asked one of the teachers to administer three released SAT items addressing the targeted content that the students had completed two weeks before as part of the GenScope Assessment Project research. However a miscommunication led the students to answer the items working in their collaborative group. This made it impossible to interpret changes across time.

other students who did not appear to understand how to solve the problem. During the feedback session, the intended activity required the group of students to negotiate a shared understanding of the correct solution in order to reach consensus that the answer was indeed worthy of full points. The students who had not participated in the construction of the correct answer in the first place appeared involved in negotiating an understanding of the correctness of that same answer. Such instances appear to be powerful and somewhat unique affordances of the intended activity.

The level of engagement in the science content overall was particularly heartening in light of the disappointing results from the GenScope Assessment Project in these same classrooms. In a very different configuration (where students were invited to review their performance assessments, but not required to) students chose to spend very little time reviewing their answers (mean time 160 seconds across 80 students) and disturbingly little of that time (only 1.6%) was coded as representing "substantive engagement" in the domain content (Hickey, Kruger, Fredrick, Schafer, & Kindfield, 2002).

Learning to collaborate. Regarding the third research question, it seemed apparent that this activity supported the learning of collaboration and collaborative inquiry. Our analysis of the discourse captured on the videos of the students attempting to score the digitized videos of their own collaboration revealed numerous successful examples of students negotiating a shared understanding of collaborative inquiry according to the six scales on the scoring rubric. For example, the videotapes showed repeatedly that the students did not initially comprehend the distinctions between the different dimensions. By the end of the group scoring effort, all of the groups appeared to have constructed a shared understanding of the difference between the various dimensions that was more robust than when they started; videotape of the whole class review session revealed further examples of clarification of some of the finer distinctions between the scales. Assuming that a more robust understanding of collaborative inquiry is fundamentally a good thing, we view this as a most positive outcome.

During the collaboration assessment activity, students clearly enjoyed watching themselves and their classmates, and appeared to develop an appreciation of the differences between the dimensions. We witnessed good-natured jibes directed at group members or classmates whose participation was especially lacking; their appropriate (but not seemingly excessive) embarrassment hinted positive consequences in subsequent collaborative activity. Obviously our research design can't confirm transfer of this new understanding. However, the teacher in one classroom asked the class if they thought watching themselves would improve future collaboration. A student responded, "Yes, because if you watch yourself, you might see things you're doing that you shouldn't be doing, and that way next time you won't be as likely to do those things." We consider it likely that the new understanding of collaborative inquiry provided these particular communities of learners with new labels and useful personal and social points of reference for negotiating effective participation in subsequent collaborative inquiry. Reflecting contemporary socio-cultural models of motivation (e.g., Hickey & McCaslin, 2001, McCaslin & Hickey 2001), we expect that our learning environment also served to attach value to these important participatory rituals, and provided points of reference for the teacher and for the students to reinforce these values.

Certainly there was plenty of room for improvements here. For example, while most groups were able to understand and score the use of evidence to make and warrant knowledge claims, many still seemed confused about the distinction between using evidence and theory (the difference between C5 and C6 in Figure 1). That the teacher found it difficult to even characterize this distinction was not surprising, given that the graduate students scoring the videotape reported difficulty differentiating them as well. We also encountered substantial technical challenges in using laptops and RealPlayer to view videos. While the audio volume of the laptops was barely loud enough for students to hear the audio, external speakers were invariably turned up too loudly.

It seems that the entire effort would have been more effective if students had been given access to benchmark examples to compare themselves to.

We also were presented with noteworthy ethical challenges regarding the recordings of student activity. For example, although all of the comments recorded during the initial collaboration activity were made in front of classmates in the first place, several comments were worthy of censorship. In one case it caused a commotion when students in one group insisted on playing one of the students vulgar comments back repeatedly.

Equity. Regarding the fourth question, we selected our two sets of classrooms so that we could examine whether the collaboration assessment was "appropriately appropriated" by very different communities of learners. It appeared that the high SES students had somewhat more experience at scientific argumentation—the teacher in that classroom indicated that the feedback model on the performance assessments was easy for her to appropriate because she "has similar conversations all the time in other contexts". As such, these classrooms seemed to more readily engage in productive discourse regarding the finer points of collaboration. While further analyses of the videotapes is needed to warrant this observations, it appeared that high SES students were more familiar with the distinction between warranted and unwarranted arguments, and thus it seemed more feasible to move into the distinction between using theory and using evidence to warrant arguments (the difference between Dimension 5 and Dimension 6). In contrast, the corresponding discourse in the lower SES classrooms settled around the differences between warranted and unwarranted arguments. In the classrooms where the teacher led the discussion of dimensions of collaboration, the teacher was observed to generate a non-scientific example to help illustrate the distinction between warranted and unwarranted arguments, and declined to elaborate on the distinction between warranting with theory and warranting with evidence.

Summary. In summary, it appears that our model of collaboration assessment is a promising means of enhancing participation in collaboration and increasing students' ability to engage in collaborative learning. Promisingly, it seems appropriate for higher and lower achieving students as well. Our model of formative collaboration assessment seems like it could be implemented on a larger scale, where its impact might be more systematically assessed. However, as implemented, our model of formative assessment of collaboration was far too labor-intensive to be implemented on a large scale. Simultaneously recording 7-8 collaborative groups was itself quite laborious; allowing those groups to then simultaneously view their videos was even more difficult.

EXTENSIONS AND COLLABORATIONS

The intensively collaborative nature of the project as described above reflects the intent of the CILT Seed Grant Program. There were additional collaborative aspects of the project not described above that are also worthy of note. One involved an IERI-funded study of early reading practices raised the issue of the quality of paired partner reading (where two beginning readers take turns reading aloud to each other). This research is being carried out by Paula Schwanenflugel at UGA, with Graduate Assistant Beth Meisinger. Paired partner reading was an important independent variable in their study and they needed an efficient way to assess it. This presented a very different curricular and research context, and thus allowed us to consider the generalizability of the assessment practice. The fact that they are scoring it live (i.e., no videotape is allowed in this district) really puts this method to the test. They initially selected four of the dimensions from the original Pomplun scale, and reworded the labels and the three-point anchors to be specific to that activity. They subsequently used four of the candidate scales from our project. The method was piloted in 2001, and they reported near perfect correspondence on the five scorings carried out.

They have indicated that they intend to use the scales during next round of classroom studies in 2002.

An additional collaborative aspect of our work was the renewed involvement of Jackie Gray and Jennifer Holbrook from Georgia Tech. Dr. Gray participated in several of the scale development meetings, and the group considered including new scales that she had developed for the continuing collaboration assessment in the LBD project. In the end, the group concluded that these dimensions to be too specific to the LBD collaboration assessment to include in the set of candidate dimensions. However, the LBD team appears to have made substantial progress in creating reliable scales that assess participation in important aspects of collaboration that we were unable to address (such as designing experimental investigations). This appears to be a promising direction when considering collaboration in more open-ended inquiry environments such as LBD.

Finally, Juan Balderas (an educational administrator in Ogden UT) coordinated a collaboration assessment and feedback activity with two teachers, Jon Contos and Vicki Winn. This ultimately proved rather challenging, in that all three were already quite occupied with existing teaching and administrative demands. The teachers reported that they had quite a bit of difficulty with videotaping equipment, and were somewhat unclear about the ultimate goal of the activity. Ms. Winn provided a detailed report that describes a worthwhile activity that was ultimately quite unlike from the efforts carried out elsewhere.

NEXT STEPS

A proposal was submitted to the National Science Foundation's ROLE (Research on Learning Environments) program in November 2001 called *Systemically Valid Science Assessment Practices for Ubiquitous Web-Based Video*. The proposal builds quite directly on the work described above, and incorporates recent technological and methodological advances. We proposed to implement the formative assessment practices described above using a variant of a tapeless real-time capture & compress/chunk & code system for event-based research known as ITMD (*Integrated Temporal Multimedia Data*). This system is currently being developed by Ken Hay and researchers at the Learning Performance Support Laboratory and elsewhere; additional information is available at http://lpsl.coe.uga.edu/VBR/. We further proposed to study the effectiveness of the formative assessment practices on student content knowledge and participation in inquiry using the CN-ARE (Constructing Networks of Action-Relevant Episodes) methodology developed by Barab and Hay (e.g., Barab, Hay, & Yamagata-Lynch, 2001; Hay & Barab, 2001).

REFERENCES

Barab, S. A., Hay, K. E., & Yamagata-Lynch, L. C. (2001). Constructing networks of activity: An in-situ research methodology. T*he Journal of The Learning Sciences 10*(1&2), 63-112.

Black, P. & Wiliam, D. (1998). Asssessment and classroom learning. Assessment in education, 5, (1).

Bransford, J. B., Brown, A. L. Cocking R. (1999) (Eds.). *How people learn: Brain, mind, experience, and school.* Committee on Learning Research and Educational Practice. Washington, DC: National Academy Press.

Driver, R., Newton, P., & Osborne, J. (2000). Establishing the norms of scientific argumentation in classrooms. *Science Education, 84,* 287-312.

Duschl, R. (2000). Making the nature of science explicit. In R. Millar, J. Osborne & J. Leach (Eds). *Improving Science Education.* pp 185-206. Philadelphia:Open University Press.

Duschl, R. A. & Gitomer, D. H. (1997). Strategies and challenges to changing the focus of assessment and instruction in science classrooms. *Educational Assessment, 4,* 37-73.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher*, 18 (9), 27-32.

Gray, J. (2001, April). *Science talk as a way to assess student transfer and learning: Implications of formative assessment.* Paper presented at the Annual Conference of the American Educational Research Association, Seattle, WA.

Graue, M. E. (1993). *Integrating theory and practice through instructional assessment. Educational Assessment, 1 (4),* 283-309.

Greeno, J. G. (1998). *The situativity of knowing, learning, & research.* American Psychologist, 53, *5-26.*

Haertel, J., & Means, B. (2000). *Stronger designs for research on educational uses of technology: Conclusions and implications.* SRI, Inc. Available online at www.sri.com/policy/designkt/found.

Hay, K. E., & Barab, S. A. (2001). Constructivism in practice: A comparison and contrast between apprenticeship and constructionist learning environments. *Journal of The Learning Sciences, 10,* 281-322.

Hickey, D. T. (1999). *Effects of* Learning by Design *on Content Knowledge, Conceptual Understanding, and Collaboration Skills in the 1998-1999 Implementation.* Unpublished Project Report. Georgia State University.

Hickey, D. T. (2000a). *Assessment, motivation, & epistemological reconciliation in a technology-supported learning environment.* Grant REC-0196225 from the National Science Foundation, Division on Research, Evaluation, & Communication to the University of Georgia.

Hickey, D. T. (2000b). *Effects of* Learning by Design *on Content Knowledge, Conceptual Understanding, and Collaboration Skills in the 1999-2000 Implementation.* Unpublished Project Report. Learning & Performance Support Laboratory, University of Georgia.

Hickey, D. T. (2001). *A pragmatic, situative framework for evaluating innovative science learning environments.* Paper presented at the annual conference of the American Educational Research Association, Seattle. (Currently under review in *Science Education*)

Hickey, D. T., Kindfield, A. C. H., Horwitz, P., & Christie, M. A. (1999). Advancing educational theory by enhancing practice in a technology supported genetics learning environment. *Journal of Education*, 181(2), 1-33.

Hickey, D. T. Kruger, A. C., Fredrick, L. D., Schafer, N. J. & Kindfield, A. C. H. (2002, August). *Balancing Formative and Summative Science Assessment Practices: Year One of The GenScope Assessment Project.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, 2001.

Hickey, D. T., & McCaslin, M (2001). Comparative and sociocultural analyses of context and motivation. In S. Volet, S. & S Järvelä (Eds.), *Motivation in learning contexts: Theoretical and methodological implications.* (pp. 33-56). Amsterdam: Pergamon/Elsevier.

Hickey, D. T., Wolfe, E. W., & Kindfield, A. C. H. (2000). Assessing learning in a technology-supported genetics environment: Evidential and consequential validity issues. *Educational Assessment, 6* (3), 155-196.

Jiménez-Aleixandre, M. P., Rodríguez, A. B., & Duschl, R. A. (2000). "Doing the lesson" or "Doing science": Argument in high school genetics. *Science Education, 84,* 757-792.

Kendall J. S., & Marzano, R. J. (1997) *Content Knowledge: A compendium of standards and benchmarks for K-12 education.* Aurora, CO: MCREL. Available on line at http://www.mcrel.org/standards-benchmarks/.

Kolodner, J. L., Crismond, D. Gray, J., Holbrook, J, Puntambekar S. (1998). Learning by design from theory to practice. In J. Kolodner, M. Guzdial, & Bruckman, A. (Eds.), *Proceedings of the 3rd International Conference of the Learning Sciences, 16-22.*

Kolodner, J. L., Gray, J.& Fasse, B. B. (in review). Promoting transfer through case-based reasoning: Rituals and practices in *Learning by Design* classrooms. *Cognitive Science Quarterly*, Vol. 1.

McCaslin, M. & Hickey, D. T. (2001). Self-regulated learning and academic achievement: A Vygotskian view. In B. Zimmerman and D. Schunk (Eds.), *Self-regulated learning and academic achievement: Theory, research, and practice, Second Edition* (pp. 227-252). Mahwah, NJ: Erlbaum .

PALS (2001). *Performance Assessment Links in Science.* A website maintained by SRI International. Accessed November 20th at http://www.ctl.sri.com/pals/ .

Pomplun, M. (1996). Cooperative groups: Alternative assessment for students with disabilities. *The Journal of Special Education, 30* (1), 1-17.

Turnstall, P, & Gipps, C. (1996). Teacher feedback to young children in formative assessment: A typology. *British Educational Research Journal, 22,* 389-404

Wenger, E. (1998). *Communities of practice: Learning, meanings, & identity.* Cambridge, England: Cambridge University Press.

Collaboration Assessment Rubric

UGA Learning & Performance Support Laboratory/Center for Interactive Learning Technology

This survey presents six dimensions that collectively cover the most important aspects collaborative inquiry. A five-point scale was constructed for each dimension. After you rate the observed collaboration on each dimension, rate your confidence in that rating. Specifically, the specific event you observed, how confident are you that you would give that same score if asked to score it again, without recalling the score assigned previously. Use 1 for *not at all confident,* 2 for *somewhat confident,* and 3 for *very confident.*

C1. *Emotional Support.* Group members help each other and show consideration for others.

1. All members are negative (physically turned away from each other, and sarcastic, most of the time.	2. Some of the members are negative and sarcastic some of the time.	3. Both negative and positive emotional support, or mostly neutral emotions, are observed	4. Some of members are providing positive comments and listening some of the time. ("Good idea")	5. All members are listening and providing support most of the time.	Collaboration Score _____ Confidence Score_____

C2. *On-Task Behavior.* Participants stay on task (and conduct self-checks on procedures if needed)

1. Group unfocused; not working on task; unclear goals and understanding of directions. (No members question or address staying on task.)	2. Group is somewhat on task some of the time. (One member questions some aspects of the procedure.)	3. Group is on task some of the time. (Some members question some aspects of the procedures)	4. Group is mostly on task most of the time. (At least one member questions some aspects of the procedure and group responds accordingly)	5. Group is on task all of the time (More than one member questions aspects of the procedure and the group responds accordingly).	Collaboration Score _____ Confidence Score_____

C3. *Conflict Management. Group* members handle conflicts (and negotiate effective solutions as needed) .

1. Conflicts prevent all progress (No negotiations of conflicts at all)	2. Conflicts interfere substantially with progress (One member of the group attempts to negotiate at appropriate opportunity	3. Conflicts interfere somewhat with progress. (Some members attempt to negotiate task at appropriate opportunities).	4. Conflicts interfere slightly with progress (Most members negotiate tasks and answers at most opportunities.)	5. Conflicts do not interfere with progress at all (all members effectively negotiate and share ideas whenever appropriate)	Collaboration Score _____ Confidence Score_____

C4 *Collaborative Participation.* Group members demonstrate collaborative participation.

1. No collaborative participation.. Work is done by individuals without assistance from others. No teaming is evident.	2. A little collaboration some of the time, or one individual dominates.	3. Some collaboration, but not all group members participate	4. A lot of collaboration, with frequent interaction between most members, most of the time.	5. Teamwork, participation and collaboration are high; decision-making is shared by all members all of the time.	Collaboration Score _____ Confidence Score_____

C5. *Evidential Knowledge Claims.* Participants make and warrant (i.e., argue for) knowledge claims based on evidence before them (after Duschl).

1. No claims. No references are made to the available evidence or data.	2. Few claims. One member refers to the available evidence or data at least once.	3. Some Claims. At least one member refers to the available evidence or data more than once.	4. Some claims and warrants. At least one member refers to the available evidence or data and uses it to make appropriate knowledge claims.	5. Many claims and warrants. All members refer to the available evidence or data and use it to make appropriate knowledge claims.	Collaboration Score _____ Confidence Score_____

C6. *Theoretical Knowledge Claims.* Participants make and warrant (i.e., argue for) knowledge claims using domain knowledge (after Duschl).

1. No claims. No references are made to relevant domain theory	2. Few claims. One member refers to relevant domain theory at least once.	3. Some Claims. At least one member refers to relevant domain theory at least once.	4. Some claims and warrants. At least one member refers relevant domain theory and uses it to make appropriate knowledge claims.	5. Many claims and warrants. All members refer to relevant domain theory to make appropriate knowledge claims.	Collaboration Score _____ Confidence Score_____

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®
TM034272

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Technology Supported Formamative and Summative Assessment of Collaborative Scientific Inquiry	
Author(s): Daniel T. Hickey, ~~[illegible]~~ Jessica Decuir, Bryon Hand Brandon Kyser	
Corporate Source: Simona Laprocina, Joy Mordica Learning + Performance Support Lab University of Georgia	Publication Date: April 2002

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1 ↑

☒

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A ↑

☐

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B ↑

☐

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature: [signature]	Printed Name/Position/Title: Daniel T. Hickey, Asst. Prof
Organization/Address: 611 Aderhold Hall Athens GA 30345	Telephone: 706 542 3157 FAX: 706 542 4321
	E-Mail Address: dhickey@coe.uga.edu Date: June 19, 2002

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
UNIVERSITY OF MARYLAND
1129 SHRIVER LAB
COLLEGE PARK, MD 20742-5701
ATTN: ACQUISITIONS

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com

EFF-088 (Rev. 2/2000)